

# RecSys Challenge 2017: Offline and Online Evaluation

Fabian Abel  
XING AG  
Hamburg, Germany  
fabian.abel@xing.com

Mehdi Elahi  
Free University of Bozen - Bolzano  
Bozen - Bolzano, Italy  
meelahi@unibz.it

Yashar Deldjoo  
Politecnico di Milano  
Milan, Italy  
yashar.deldjoo@polimi.it

Daniel Kohlsdorf  
XING AG  
Hamburg, Germany  
daniel.kohlsdorf@xing.com

## ABSTRACT

The ACM Recommender Systems Challenge 2017<sup>1</sup> focused on the problem of job recommendations: given a new job advertisement, the goal was to identify those users who are both (a) interested in getting notified about the job advertisement, and (b) appropriate candidates for the given job. Participating teams had to balance between user interests and requirements for the given job as well as dealing with the cold-start situation. For the first time in the history of the conference, the RecSys challenge offered an online evaluation: teams first had to compete as part of a traditional offline evaluation and the top 25 teams were then invited to evaluate their algorithms in an online setting, where they could submit recommendations to real users. Overall, 262 teams registered for the challenge, 103 teams actively participated and submitted together more than 6100 solutions as part of the offline evaluation. Finally, 18 teams participated and rolled out recommendations to more than 900,000 users on XING<sup>2</sup>.

## CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Data mining*; *Test collections*;

## KEYWORDS

Recommender Systems; Challenge; Cold Start

## 1 INTRODUCTION

The ACM Recommender Systems Challenge is an annual competition that focuses on the design and development of the best performing recommendation algorithms for a particular scenario. From its inception in 2010 until the eighth iteration in 2017, the challenge has experienced a substantial gain in interest and has drawn a lot of participants from industry and academia. It thus became a key part of the ACM RecSys Conference series [3].

<sup>1</sup><http://2017.recsyschallenge.com>

<sup>2</sup><https://xing.com>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

RecSys'17, , August 27–31, 2017, Como, Italy.

© 2017 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4652-8/17/08.

<https://doi.org/10.1145/3109859.3109954>

The 2017 edition of the challenge is organized jointly by XING, Politecnico di Milano<sup>3</sup> and Free University of Bozen - Bolzano<sup>4</sup>. This edition focused on a reciprocal recommendation problem: recommending job advertisements to people who are interested in the job and who are at the same time also good candidates for the job role that needs to be filled in the corresponding company. This problem had to be tackled in the cold-start situation [2].

XING, as an industrial organizer of the challenge, is a career-oriented social networking site, with around 20 Million users and around one Million open job offers on the platform. For the challenge, we published large training datasets of anonymized user profiles, job postings, and interactions between them.

The RecSys challenge 2017 started with an offline evaluation round. The top 25 teams qualified for the subsequent online evaluation round during which the participating teams were actually pushing recommendations to real XING users.

## 2 DATASET

For both offline and online evaluation, we provided a dataset that features the following details<sup>5</sup>:

**interactions:** all transactions between a user and an item such as impressions, clicks, bookmarks, deletions and recruiter interests (= recruiter interacted with the candidate).

**users:** details about users such as job title, skills, career level, location, estimated willingness to change jobs, etc.

**items:** details about the job advertisement including fields such as job title, required skills, location, etc.

**target items:** list of item IDs which the teams were allowed to push to interested users and for which the teams were supposed to identify matching candidates.

**target users:** set of user IDs for which the teams were allowed to recommend items and which were allowed to appear as candidates.

During the online evaluation (see Section 4), the list of target items and target users changed on a daily basis. The training datasets were enriched with noise and anonymized in the same way as during the RecSys Challenge 2016 [1]. Basic statistics about the training datasets, that were handed out for the offline evaluation and shortly before the online evaluation, are listed in Table 1.

<sup>3</sup><https://www.polimi.it>

<sup>4</sup><https://www.unibz.it>

<sup>5</sup><http://2017.recsyschallenge.com/#dataset>

**Table 1: Statistics about datasets, used in offline and online evaluations.**

	Offline	Online
users	1,497,020	963,346
items	1,306,054	853,058
impressions	314,501,101	88,719,060
clicks	6,867,579	3,552,973
bookmarks	281,672	184,061
replies	117,843	96,907
deletes	906,836	360,645
recruiter interests	100,971	35,715

### 3 CHALLENGES

During RecSys Challenge 2016, teams had to focus solely on estimating how relevant a job is for a given user [1]. In 2017, teams had to tackle multiple challenges at once.

*Balancing user interest and recruiter demands:* job recommendations should be relevant to the users. However, at the same time the users who receive those job recommendations also need to be appropriate candidates for the given job, i.e. the corresponding recruiter should show an interest in the candidate.

*Balancing relevance and revenue:* for some of the job advertisements, companies actually paid money to publish the ad on XING. Moreover, some users pay for premium subscriptions. Teams had to balance between relevance of recommendations and monetary aspects, i.e. the money that is earned with a successful recommendation.

*Novelty / sparsity:* recommendations had to be computed particularly for newly created job postings (new items). At the time a posting was supposed to be recommended to users, most of the postings did not receive any interaction yet.

*Smart targeting of recommendations:* also, the participating teams had to estimate how likely it is that a user is actually interested in job recommendations, e.g. users that are not interested in job recommendations may delete recommendations or disable notifications about recommendations in case they receive too many. The latter was primarily relevant for the teams that participated in the online challenge.

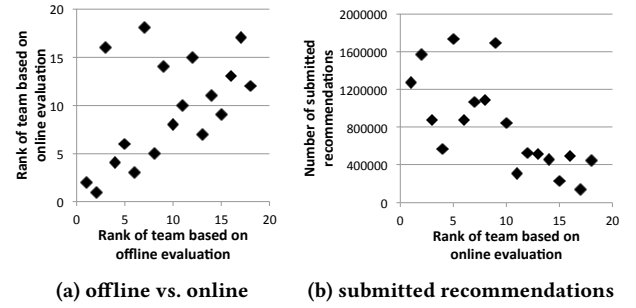
### 4 EVALUATION

The above challenges were also reflected in the evaluation metrics<sup>6</sup>. For example, recommendations for which the candidate and recruiter showed mutual interest led to a higher score compared to recommendations that just triggered the candidate to click on the job advertisement. Furthermore, successfully recommending paid content to premium users also led to higher scores. The same evaluation metrics is used in both offline evaluation phase and the subsequent online evaluation phase:

**offline evaluation:** during this phase which started on March 3rd 2017 and ended on April 16th, teams trained their algorithms on a fixed historic dataset and had to submit their solutions for a fixed set of target items and users (see Table 1).

**online evaluation:** this part of the challenge started on May 1st 2017 and ended on June 4th. Participating teams were

<sup>6</sup><http://2017.recsyschallenge.com/#evaluation>



**Figure 1: Correlation between: (a) ranks achieved during the offline and online evaluations and (b) ranks and the number of submitted recommendations (online)**

equipped with an up-to-date training dataset before the start of the online challenge to prepare and re-train their algorithms and models. New target job advertisements, new interactions, etc. were then handed out on a daily basis. The teams had 24 hours time to submit their recommendations which were then actually rolled out on XING (Web, mobile apps and emails). The evaluation metric was applied on the interaction data that was collected within seven days after the recommendations were submitted<sup>7</sup>.

### 5 RESULTS AND CONCLUSION

Figure 1a compares the performance of the teams during the offline evaluation and online evaluation: for most teams, we observe that they managed to confirm their offline evaluation performance also during the online challenge. Only for a few teams, the performance dropped strongly, e.g. one team lost more than 10 positions in the ranking. The recommendations contributed by that team also showed the lowest specificity regarding user targeting: one item was, on average, recommended, to 86 users. Other teams typically recommended the same item to 4-11 users.

We also see that the pure amount of recommendations that were submitted during the online challenge did not decide about the winner (see Figure 1b), for example: the team that submitted the highest number of recommendations ended up at rank 5.

In summary, we saw that there was high engagement among the participants: 103 teams were active during the challenge. As part of the offline evaluation, each team submitted, on average, 59.4 different solutions. During the online evaluation, the qualified teams successfully proved that their algorithms also work in a live system. Overall, the participants managed to rolled out more than 14.7 Million recommendations to more than 900,000 XING users.

### REFERENCES

- [1] Fabian Abel, András Benczúr, Daniel Kohlsdorf, Martha Larson and Róbert Pálóvics. 2016. RecSys Challenge 2016: Job Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 425–426.
- [2] Mehdi Elahi, Francesco Ricci, and Neil Rubens. 2016. A survey of active learning in collaborative filtering recommender systems. In *Computer Science Review* 20, 29–50.
- [3] Alan Said. 2016. A Short History of the RecSys Challenge. In *AI Magazine*, 37, 4.

<sup>7</sup><https://github.com/recsyschallenge/2017/tree/master/online-evaluation>